A robustness approach to theory building: A case study of language evolution

Liz Irvine (elizabeth.irvine@cin.uni-tuebingen.de)
Philosophy of Neuroscience, Werner Reichardt Centre for Integrative Neuroscience,
Otfried-Müller-Str. 25, 72076 Tübingen, Germany

Seán G. Roberts (sean.roberts@mpi.nl)
Language and Cognition Department, Max Planck Instutite for Psycholinguistics
Wundtlaan 1, Nijmegen 6525 XD The Netherlands

Simon Kirby (simon@ling.ed.ac.uk)
Language Evolution and Computation Research Unit, School of Philosophy, Psychology
Language Sciences, University of Edinburgh, Dugald Stewart Building, 3 Charles Street, Edinburgh, EH8 9AD, UK

Abstract

Models of cognitive processes often include simplifications, idealisations, and fictionalisations, so how should we learn about cognitive processes from such models? Particularly in cognitive science, when many features of the target system are unknown, it is not always clear which simplifications, idealisations, and so on, are appropriate for a research question, and which are highly misleading. Here we use a case-study from studies of language evolution, and ideas from philosophy of science, to illustrate a robustness approach to learning from models. Robust properties are those that arise across a range of models, simulations and experiments, and can be used to identify key causal structures in the models, and the phenomenon, under investigation. For example, in studies of language evolution, the emergence of compositional structure is a robust property across models, simulations and experiments of cultural transmission, but only under pressures for learnability and expressivity. This arguably illustrates the principles underlying real cases of language evolution. We provide an outline of the robustness approach, including its limitations, and suggest that this methodology can be productively used throughout cognitive science. Perhaps of most importance, it suggests that different modelling frameworks should be used as tools to identify the abstract properties of a system, rather than being definitive expressions of theories. **Keywords:** Language Evolution; Cultural Evolution; Robustness

Introduction

A central question in the field of the evolution of language is whether linguistic structure is mainly a product of domain-specific genetic constraints, or of cultural transmission. However, the cultural evolution of language is difficult to study because there is little direct evidence available. Simulations[1] make it possible to study of the dynamics of cultural evolution, but often include highly simplified mechanisms of learning. Human experiments obviously use a realistic learning mechanism, but present the problem that test subjects already know natural languages, and it is difficult to control for individual differences in learning. While recent work suggests that compositional linguistic structure emerges in iterated learning contexts under pressures to be learnable and expressive (Kirby, Cornish, & Smith, 2008), problems with simulations and human experiments potentially make it difficult to justify theoretical claims about the evolution of language.

We explore how the notion of 'robustness' from philosophy of science (e.g. Wimsatt, 1981; Weisberg & Reisman, 2008) can be used to support claims in language evolution; specifically, we argue that emergent compositional structure is a robust property of iterated learning. We suggest how to extend this methodology to similar claims across cognitive science, that are based on unrealistic models, and little direct empirical evidence.

Language Evolution

Before discussing the notion of robustness, it is first necessary to introduce work on the iterated learning model (ILM) in language evolution as this will be used as a case-study. The ILM looks at how culturally transmitted systems (e.g. a mapping between linguistic signals and meanings) change by being repeatedly transmitted through a bottleneck (Kirby, 2000). The bottleneck is a restriction of information that could be due to a finite limit on the information to be transferred or a restricted set of meanings to be described. The bottleneck causes the system to change over time, usually towards a more compressible relationship between signals and meanings. This can be interpreted as a pressure on the language to become more 'learnable' by the next generation. In the extreme case, the variation in the system reduces so that there is only one signal. Opposing this is a pressure for expressivity (e.g. a need to distinguish between meanings). A perfectly expressive linguistic system has a different signal for each meaning.

Smith, Kirby, and Brighton (2003) showed that an optimal solution under these two pressures is compositionality: the meaning of a signal is composed of sub-meanings expressed by sub-strings of the signal. This means that there are fewer signal components to be learned than individual meanings, and the signals of unobserved meanings can be re-constructed accurately. The demonstration that cultural transmission can lead to complex linguistic structure contrasts with theories that see linguistic structure as primarily deriving from innate, domain-specific constraints (Chomsky, 1965, see Kirby,

[1] In this paper we maintain a distinction between 'models', which are analytically analysable descriptions of a system and 'simulations' which are individual numerical implementations of a model. While some results are derived analytically from models, others come from numerical simulation. Iterated learning experiments with human subjects can also be seen as simulations of the process of language evolution.

Dowman, & Griffiths, 2007).

The initial work on the iterated learning model involved computational simulations. Instead of committing to a particular model or simulation framework, a range of computational techniques were used as tools to demonstrate the principles of the iterated learning model. These include grammar induction models (Kirby, 2000), exemplar models (Batali, 2002), neural network models (Kirby & Hurford, 2002; Swarup & Gasser, 2008) and self-organising maps (Worgan & Damper, 2008).

The next step involved translating the ILM into a laboratory experiment. Kirby, Cornish, and Smith (2008) demonstrated that the emergence of compositional structure could be observed in an artificial language which was learned, produced, transmitted and then learned again by human subjects. Participants were exposed to pairings of nonsense words and meanings and asked to memorise them. The meanings were images with structured semantic dimensions: shape (circle, triangle, square), colour (red, blue, black) and movement (horizontal, bouncing, or spiraling motion). Participants were trained on a sub-set of the whole meaning space, but they were then asked to produce a label for every meaning. The labels that were produced became the training data for the next participant. This meant that the language changed as it was transmitted from participant to participant, mirroring the cultural transmission of language.

By this process, the language adapted to two pressures. A bottleneck on transmission was present, because the participants were not trained on all labels. This put a pressure on the language to become more faithfully transmitted. With only this pressure, the number of distinct labels in the language declined. These languages were easy to learn, but not expressive. To counter this, a pressure for expressivity was added by excluding homonyms from the training subset. Under both pressures, the language adapted to become learnable and expressive by becoming compositional. That is, instead of labels being distinct and holistic, sub-parts of each label consistently referred to sub-parts of each meaning. For instance, in one emergent language, all meanings which included the colour blue began with an 'L' while all meanings that included the spiralling movement ended with 'PILU' (see Kirby, Cornish, & Smith, 2008, p. 10684). This meant that the language was both easy to learn and could express all meanings distinctly.

The ILM experiment showed that compositional structure could emerge spontaneously due to the process of cultural transmission. The results mirrored those of the computational simulations, leading to the experiments being thought of as simulations with human participants (see Kirby, Smith, & Cornish, 2008).

The ILM sparked a lineage of experimental simulations testing different constraints and assumptions of the original simulation, including replacing the exclusion of homonyms with a pressure for communication between two participants (Matthews, Kirby, & Cornish, 2010; Silvey, Kirby, & Smith, 2012; Navarro & Perfors, 2011; Tamariz, Cornish, Roberts, & Kirby, 2012; Verhoef & Boer, 2012). Bringing the process full circle, principles elucidated through the human simulations motivated new computational simulations (Smith, Tamariz, Cornish, & Kirby, 2013). There are differences between these studies, for example the precise distribution of letters in the strings that emerge is not robust across computational and human simulations since human distinctions between vowels and consonants were not built into the computational agents. However, in each case the results were compatible with the theory of structure emerging though repeated transmission through a bottleneck under pressures for learnability and expressivity.

Problems with abstraction and transparency

Computational models have many advantages: the internal states of individuals are transparent and quantifiable; the exact amount of noise is quantifiable; and exploring the parameter space can be easier than running alternate conditions in human experiments (e.g. Reali & Griffiths, 2010 model an infinite population). However, the abstractions inherent in computational models can be a weakness because a model's implications rely on the ability to translate between the abstractions and the real world. Computational models of the cultural evolution of language may simplify the representation of linguistic units, learning processes or psychological mechanisms. Simplifying assumptions might be made such as agents sharing an innate, conceptual space for words and meanings (Vogt, 2005) or being able to observe intended meanings (Worgan & Damper, 2008).

In contrast, laboratory experiments with real humans include artificial languages with concrete analogues to real languages and real learning mechanisms. However, while the learning mechanisms are realistic, they are opaque. It is difficult to deduce the precise mental processes that lead to the emergence of structure. Because of this complexity, it is difficult to maintain absolute experimental control. Furthermore, the participants already have full knowledge of a compositional language. This is a potential confound since the emergent structure may just be a reflection of the participants' existing language rather than being caused by the same process that proto-linguistic humans underwent (e.g. Flynn, 2008; Chomsky, 2011). However, as we shall see below, these problems can be addressed by demonstrating that the outcome is a robust property across different models..

Aims of models and simulations

One crucial question in this area of research is what simulations are used to learn about, and thus what kind of theoretical inferences can possibly be warranted from them. Simulations of cultural transmission are not intended as instantiations of human language learning, nor the evolution of a real language. However, the ILM is informative about systems that are transmitted through a bottleneck, and that become more structured as a result. The simulations, then, are an example

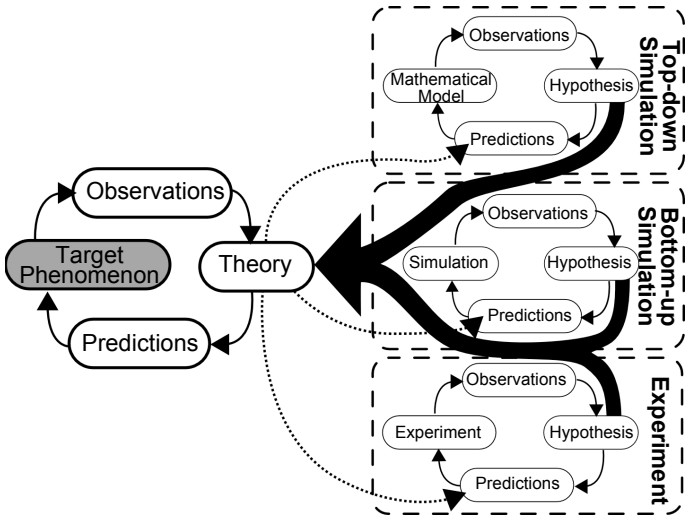

Figure 1: How simulations, models and lab experiments can contribute to theories of language evolution. Potential problems in the translation from the target phenomenon to models, simulations, and experiments (dotted arrows) are spread over many approaches, while the discovery of robust properties found across all these approaches provide support for the central theory (thick arrows).

of cultural transmission's effect on structure in cultural phenomena, where human language is one of these more general phenomena. Results from simulations of iterated learning can therefore inform the general theory of iterated learning[2].

In the field of language evolution, simulations of iterated learning are used as thought experiments to show that nativist assumptions are not necessary to account for the emergence of complex linguistic structure. Computational simulations can be more powerful than traditional thought experiments because they must usually be instantiated at a more technical level, and because they allow complex interactions and structures which might be unintuitive (Bedau, 1998).

Di Paolo, Noble, and Bullock (2000) suggest a workable methodology for opaque thought experiments:

1) Exploratory phase: Explore model behaviour, observe patterns

2) Experimental phase: Formulate hypotheses that organise patterns ('explanatory organisation'). Some patterns will be explained by the model dynamics directly. Some patterns will be explained through other observed patterns ('indirect explanation').

3) Explanatory phase: Relate organisation of observations to the theories about natural phenomena, explain the consequences.

[2]However, since the target of study is the cultural transmission process (how systems change by being repeatedly transmitted through a bottleneck) both computational and human simulations can arguably be seen as *actual instantiations* of this highly abstract generic phenomenon.

This idea is very similar to recent work in philosophy of science (particularly philosophy of economics), where models and simulations can be seen as 'credible worlds', constructed to explore general theoretical principles (Sugden, 2000, 2009). Recent work on model-based theorising (Weisberg, 2007; Godfrey-Smith, 2006), largely based on examples from population biology, incorporates similar stages of modelling, going through stages of model construction, model analysis (stage 1 and 2) and the (optional) stage of the exploration of how well the model 'fits' the target system, and thus which general principles can be learned from the model (stage 3). Here too, there are questions about how these constructed, and often highly simplified worlds relate to real world systems.

There are two main ways model-world fit can be evaluated, both found in cognitive science. One is to consider the translation from the target system to the model/simulation, for example when models and simulations are used as abstract analogues of a concrete phenomenon (e.g. eye saccades during reading). Here knowledge about the target system is used to construct the model (though of course it may involve simplification, idealisation and so on). Model-world fit is analysed by considering how different the model is from our knowledge of the target system (perhaps it fails to capture important structural or causal features).

The other way to consider model-world fit is to consider the converse; the translation from a model/simulation back to a the target system. Here, a model-simulation is constructed using relevant background knowledge of a family of targets, but not intended to represent any particular target system. Once the principles governing the model/simulation are established, the researcher then looks to see if the model/simulation actually captures any targets in the real world. If any are found, then the researcher infers that the same principles govern the model/simulation and these target systems. It is this kind of translation that is found in Weisberg and Godfrey-Smith's description of model-based science, where justifications must be given for inferences from properties found in a model to properties of a target system. In language evolution, this inference goes from the results of iterated learning paradigms to real cases of language evolution.

In order to support this inference, there are a number of ways that model-target fit can be evaluated. The role that the notion of robustness plays in these evaluations is discussed below, and linked to other cases in cognitive science.

Model-Target Fit and Robustness

Robust properties

Robust properties are those that are consistently found across a set of different models, suggesting that it is an 'important' property that derives not from incidental features of the model (e.g. its particular assumptions and simplifications), but from the core structure found across all the models (Levins, 1966; Wimsatt, 1981; Weisberg, 2006; Weisberg & Reisman, 2008).

As Levins originally put it, "if these models, despite their different assumptions, lead to similar results, we have what we can call a robust theorem that is relatively free of the details of the model. Hence, our truth is the intersection of independent lies." (Levins, 1966, p. 20). As detailed above, in the case of language evolution the emergence of linguistic structure is a robust property of iterated learning systems under the pressures of expressivity and learability.

However, Weisberg and Reisman (2008) identify several (related) kinds of robustness: parameter, structural and representational robustness. First, a model can be robust (i.e. give roughly the same results) for a wide range of parameter settings. Second, a set of crucial causal components that make up the 'core structure' noted above give rise to structural robustness. Finally, representational robustness refers to the range of model descriptions (e.g. programming languages or mathematical formalisms used) for which a model still gives rise to the same results. Of most relevance in the sections below are structural and representational robustness, detailed below.

Multi-model method

One way that robustness analyses figure in language evolution research is through the use of different types of models. Researchers in the field of cultural evolution see models as tools, not necessarily as reflections of theories (e.g. Cornish, Tamariz, & Kirby, 2009). As tools, they need not commit the researcher to particular methodological approaches (agent-based or mathematical) nor particular theories of cognition (e.g. humans as Bayesian learners or frequentist learners).

Testing the same ideas across a range of mathematical models and computational simulations, but also across different formalisms or experimental setups within each broad method, is a standard way to explore robust properties. This corresponds to both structural and representational robustness noted above. The same core structures are found across these models, but other variables can differ (e.g. learning algorithm, size of population). That these additional variations do not affect the core finding (emergence of linguistic structure) provides support for the claim that the core structures really are the essential causal components in these models. Further, that these models can be constructed in a range of computational and mathematical frameworks, and still give rise to emergence linguistic structure, means that these results are not related to specific features of these frameworks - they are representationally robust. Both provide support for the claim that linguistic structure is a robust property of ILM models.

However, this method of constructing and comparing a range of computational and mathematical models is not widely found in current practice in cognitive science, where researchers are often wedded to particular frameworks. The field of language evolution may require this kind of approach because its precise object of study is still being identified, so the key variables and relations to consider are not yet obvious. For example, it is difficult to intuit about the abstract properties of a culturally transmitted linguistic system underpinned by genetic constraints (Christiansen & Kirby, 2003).

Human simulations

Another crucial way of exploring the relations between simulations and real systems, used in language evolution, is to replace computer agents in agent-based models with human subjects (Kirby, Cornish, & Smith, 2008). The inclusion of this much wider range of structural features (such as real biological learning mechanisms) provides a strong test for claims about the core structural features and representational robustness of ILM models. However, that subjects already know natural languages is seen as a strong confounding factor in the interpretation of human simulations. There are two standard responses to this.

Firstly, there are the experimental controls. Compositional structure does not always emerge in these simulations, only when both the pressures of expressivity and learnability are applied. Also, the human participants, far from deliberately introducing familiar linguistic structures, rarely expressed an understanding of the principles behind the experiment, most even not noticing that they were being tested on meanings that they had not been taught (Kirby, Cornish, & Smith, 2008).

Secondly, there is an argument based on the notion of robustness. Human simulations can be seen as further explorations of structural and representational robustness, that include both actual biological mechanisms (e.g. learning), but also potentially problematic factors (subjects already know natural languages). That linguistic structure still reliably emerges from iterated learning paradigms under pressures for learnability and expressivity, even when significant variables are changed, provides more evidence that the emergence of linguistic structure is a robust property of these models.

Summary: Learning from simulations in the ILM

Since there is little empirical evidence about the facts of language evolution, the strength of model-target 'fit' may not be most convincingly based on the comparison to the real world, but on the robust properties found under various simplifications and idealisations of real world target systems. Even if we are not sure of how precisely to represent a target system, the fact that many highly idealised representations of the same system make a similar prediction (e.g. emergence of compositionality), can be sufficient to suggest that this is what is happening in the target system itself. In this case, theoretical claims based on robust properties already have some degree of 'fit' with target systems, purely because of the nature of robustness.

Robustness in Cognitive Science

The sections above illustrate how to productively use robustness notions in cognitive science. Modelling the same process over different formalisms or frameworks, and over computational, mathematical and human models and simulations, can help identify general principles and core variables and constraints. Each particular model need not have high model-world fit, but together the emergence of a robust property cuts

through these problems. However, there are of course limits to robustness approaches, and, as with any other methodology, there are no clear cut rules about its application.

First, it is not clear, in general, how many independent lines of evidence (i.e. different models/simulations) one must have in order to identify a 'real' robust property, and the core causal structure that gives rise to it. Yet this may become more clear in specific contexts. In some cases two or three very different models/simulations (e.g. containing very different assumptions) might be sufficient to warrant an inference to the existence of a robust property. Alternatively, a larger group of similar models/simulations that together survey a wide range of alternative assumptions may be required. The identification of surprising convergence across different models will always be context dependent.

Second, robustness analyses can be misleading. One might identify a robust property, and the causal structure that gives rise to it, on the basis of different models that all incorporate the same erroneous assumptions. An inference that this causal structure is also found in the world, and explains some cognitive phenomenon, would therefore be unwarranted. One might also make mistakes in identifying the robust property (perhaps it is more or less specific than found in the models), and what the relevant causal structure is.

Clearly, robustness approaches are defeasible, just as any other methods are. Yet the promotion of the use of a wide range of different frameworks found in robustness-based approaches may minimise the kind of errors identified above, or identify them earlier than approaches that stay within one modelling framework. Further, models will still be held accountable to the usual range of relevant empirical and theoretical work. Therefore robustness analyses should be seen as a additional methodological tool that can help to test and strengthen theoretical claims that are largely made on the basis of models and simulations.

Finally, one might question whether robustness analyses can not only support theoretical claims (as illustrated above), but also show when they are unfounded. In fact, it seems that criticisms of overfitting, highly parameterised models (e.g. Pitt & Myung, 2002), often based on model comparisons that include controls for model complexity (e.g. Hansen & Yu, 2001), do just this. Low-parameter models with a stable core of causal components tend to be favoured in cognitive science, which is entirely consistent with a preference for high levels of parameter and structural robustness.

One implication of the use of robustness analyses is that traditional debates about the validity of different modelling approaches may not be constructive. For example, researchers have debated whether 'bottom-up' or 'top-down' approaches are the most productive for researching cognition (Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010; McClelland et al., 2010). With a robustness approach, the question is not about which provides more realistic models or which can provide clearer analytic results, but how they can complement each other's strengths and weaknesses. In this case, it makes sense for researchers to use both approaches to identify robust properties, and thus converge on mutually supported theories.

Conclusion

This paper used research in language evolution to illustrate a robustness approach to modelling in cognitive science. It showed how robustness analyses support the identification of linguistic structure as robust property of the processes of cultural transmission, as modelled and simulated across a range of mathematical models, computational simulation frameworks and human experiments.

The robustness approach outlined here strongly contrasts with typical practices in cognitive science, where the aim is often to develop a single model, developed in a specific modelling framework, to account for a narrow range of data. Often, though not always, such practices generate models that lack predictive power and generality, and parameter and structural robustness. With the realisation that such models may have little to do with actual cognitive processes, pressures from new statistical methods of model comparisons are starting to force alternative methods of model construction.

A robustness approach directly promotes the development of models with high parameter, structural and representational robustness. These are often seen as positive features of models, as the output of such models can be traced to the activities of a set of core causal components, not to specific parameter settings, or to artifactual features of the modelling framework used. The use of multiple modelling/simulation frameworks also makes it easier to identify artefacts and core variables in models. Finally, robustness analyses offer a way of providing support for theoretical claims when there is little direct empirical evidence available. In this case, a robustness approach stands as a powerful alternative approach to modelling in cognitive science, and one we recommend highly.

Acknowledgments

Many thanks to Hannah Cornish for useful discussions.

References

Batali, J. (2002). The negotiation and acquisition of recursive grammars as a result of competition among exemplars. In T. Briscoe (Ed.), *Linguistic evolution through language acquisition: Formal and computational models* (chap. 5). Cambridge University Press.

Bedau, M. (1998). Philosophical content and method of artificial life. In T. Bynum & M. J. H. (Eds.), *The digital phoenix: How computers are changing philosophy* (p. 135-152). Basil Blackwell, Oxford.

Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press (MA).

Chomsky, N. (2011). On the poverty of the stimulus. In *Talk given at UCL division of psychology and language sciences, 11/10/2011*.

Christiansen, M., & Kirby, S. (2003). Language evolution: The hardest problem in science? *Studies in the evolution of language*, *3*, 1–15.

Cornish, H., Tamariz, M., & Kirby, S. (2009). Complex adaptive systems and the origins of adaptive structure: what experiments can tell us. *Language Learning*, *59*, 187–205.

Di Paolo, E. A., Noble, J., & Bullock, S. (2000). Simulation models as opaque thought experiments. In M. A. Bedau, J. S. McCaskill, N. Packard, & S. Rasmussen (Eds.), *Seventh international conference on artificial life* (p. 497-506). MIT Press, Cambridge, MA.

Flynn, E. (2008). Investigating children as cultural magnets: do young children transmit redundant information along diffusion chains? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1509), 3541.

Godfrey-Smith, P. (2006). The strategy of model-based science. *Biology and Philosophy*, *21*(5), 725–740.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. (2010). Probabilistic models of cognition: exploring representations and inductive biases. *Trends in Cognitive Sciences*, *14*(8), 357–364.

Hansen, M. H., & Yu, B. (2001). Model selection and the principle of minimum description length. *Journal of the American Statistical Association*, *96*(454), 746-774.

Kirby, S. (2000). Syntax without natural selection: How compositionality emerges from vocabulary in a population of learners. In C. Knight (Ed.), *The evolutionary emergence of language: Social function and the origins of linguistic form* (pp. 303–323). Cambridge University Press.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *PNAS*, *105*(31), 10681-10686.

Kirby, S., Dowman, M., & Griffiths, T. L. (2007). Innateness and culture in the evolution of language. *PNAS*, *104*(12), 5241-5245.

Kirby, S., & Hurford, J. (2002). The emergence of linguistic structure: An overview of the iterated learning model. In A. Cangelosi & D. Parisi (Eds.), *Simulating the evolution of language* (pp. 121–148). Springer Verlag, London.

Kirby, S., Smith, K., & Cornish, H. (2008). Language, Learning and Cultural Evolution: How linguistic transmission leads to cumulative adaptation. In R. Cooper & R. Kempson (Eds.), *Language in flux: Dialogue coordination, language variation, change and evolution.* London: College Publications.

Levins, R. (1966). The strategy of model building in population biology. *American Scientist*, 421–431.

Matthews, C., Kirby, S., & Cornish, H. (2010). The cultural evolution of language in a world of continuous meanings. In A. Smith, M. Schouwstra, B. de Boer, & K. Smith (Eds.), *The evolution of language: Proceedings of evolang 2010.*

McClelland, J., Botvinick, M., Noelle, D., Plaut, D., Rogers, T., Seidenberg, M., et al. (2010). Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences*, *14*(8), 348–356.

Navarro, D. J., & Perfors, A. F. (2011, Jan). Hypothesis generation, sparse categories, and the positive test strategy. *Psychol Rev*, *118*(1), 120-34.

Pitt, M. A., & Myung, J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences*, *6*, 421-425.

Reali, F., & Griffiths, T. (2010). Words as alleles: connecting language evolution with bayesian learners to models of genetic drift. *Proceedings of the Royal Society B: Biological Sciences*, *277*(1680), 429–436.

Silvey, C., Kirby, S., & Smith, K. (2012). The coevolution of words and meanings. In *Proceedings of the european human behaviour and evolution association confrence, 2012.*

Smith, K., Kirby, S., & Brighton, H. (2003). Iterated learning: a framework for the emergence of language. *Artificial Life*, *9*(4), 371-386.

Smith, K., Tamariz, M., Cornish, H., & Kirby, S. (2013). Language structure is a trade-off between compression and expression. In *Proceedings of the 35th annual cognitive science society conference (cogsci13).*

Sugden, R. (2000). Credible worlds: the status of theoretical models in economics. *Journal of Economic Methodology*, *7*(1), 1–31.

Sugden, R. (2009). Credible worlds, capacities and mechanisms. *Erkenntnis*, *70*(1), 3–27.

Swarup, S., & Gasser, L. (2008). Simple, but not too simple: Learnability vs. functionality in language evolution. In *The seventh evolution of language conference.*

Tamariz, M., Cornish, H., Roberts, S., & Kirby, S. (2012). The effect of generation turnover and intelocutor negotiation on linguistic structure. In T. C. Scott-Phillips et al. (Ed.), *The Evolution of Language: Proceedings of the 9th International Conference* (p. 555-556). World Scientific.

Verhoef, T., & Boer, B. de. (2012). Holistic or synthetic protolanguage: Evidence from iterated learning of whistled signals. In T. C. Scott-Phillips, M. Tamariz, E. A. Cartmill, & J. R. Hurford (Eds.), *The evolution of language: Proceedings of the 9th international conference (evolang9)* (p. 368-375). World Scientific.

Vogt, P. (2005). Meaning development versus predefined meanings in language evolution models. In *In packkaelbling* (pp. 1154–1159).

Weisberg, M. (2006). Robustness analysis. *Philosophy of Science*, *73*(5), 730–742.

Weisberg, M. (2007). Who is a modeler? *The British journal for the philosophy of science*, *58*(2), 207–233.

Weisberg, M., & Reisman, K. (2008). The robust volterra principle. *Philosophy of science*, *75*(1), 106–131.

Wimsatt, W. (1981). Robustness, reliability, and overdetermination. *Scientific inquiry and the social sciences*, 124–163.

Worgan, S. F., & Damper, R. I. (2008). Removing 'mind-reading' from the iterated learning model. In A.D.M. Smith et al. (Ed.), *The evolution of language: Proceedings of the 7th international conference (EVOLANG7), Barcelona.* World Scientific Publishing Co.